Turning a blind eye, but not the other cheek: On the robustness of costly punishment

Kriss, Peter H ; Weber, Roberto A ; Xiao, Erte

Abstract: The willingness to punish norm violation is an important component of many legal and social institutions, and much prior research demonstrates an apparent willingness to incur costs to punish individuals who act unfairly. But, will people rely on "excuses" to get out of having to act on costly punishment intentions, as they do with other costly pro-social acts? And how may the answer to this question depend on whether the punisher is the victim of a norm violation or an independent third party? We conduct an experiment and find that third parties punish reluctantly: although they indicate a preference to punish, they choose to avoid the opportunity to punish when they can do so without explicitly revealing that this is their preference. In contrast, second parties, who have been directly wronged, are resolute punishers—they actively seek out the opportunity to punish, even misrepresenting random outcomes in order to ensure that punishment is implemented. Our findings highlight important differences in the motives underlying second- and third-party punishment.

DOI: https://doi.org/10.1016/j.jebo.2016.05.017

Originally published at:
Kriss, Peter H; Weber, Roberto A; Xiao, Erte (2016). Turning a blind eye, but not the other cheek: On the robustness of costly punishment. Journal of Economic Behavior Organization, 128:159-177.
DOI: https://doi.org/10.1016/j.jebo.2016.05.017

Turning a Blind Eye, But Not the Other Cheek: On the Robustness of Costly Punishment

Peter H. Kriss
Medallia, Inc., Palo Alto, CA 94306, USA
peterkriss@alumni.cmu.edu

Roberto A. Weber
Department of Economics, University of Zurich, CH-8006 Zurich, Switzerland
roberto.weber@econ.uzh.ch

Erte Xiao
Department of Economics, Monash University, Clayton, VIC 3800, Australia
erte.xiao@monash.edu

Abstract: The willingness to punish norm violation is an important component of many legal and social institutions, and much prior research demonstrates an apparent willingness to incur costs to punish individuals who act unfairly. But, will people rely on "excuses" to get out of having to act on costly punishment intentions, as they do with other costly pro-social acts? And how may the answer to this question depend on whether the punisher is the victim of a norm violation or an independent third party? We conduct an experiment and find that third parties punish reluctantly: although they indicate a preference to punish, they choose to avoid the opportunity to punish when they can do so without explicitly revealing that this is their preference. In contrast, second parties, who have been directly wronged, are resolute punishers—they actively seek out the opportunity to punish, even misrepresenting random outcomes in order to ensure that punishment is implemented. Our findings highlight important differences in the motives underlying second- and third-party punishment.

JEL Classifications: C72, C92, D64

Keywords: experiment, third-party punishment, second-party punishment, fairness

Acknowledgement We thank participants at 2011 North-American Meeting of the Economic Science Association and 2012 International Meetings of the Economic Science Association for valuable comments. We are grateful to Ernst Fehr, Holger Herz, Björn Bartling, Frederic Schneider, and Donja Darai for helpful suggestions. We gratefully acknowledge the National Science Foundation (SES-0961341) for funding that supported this research.

1. Introduction

Numerous studies provide evidence that individuals' willingness to incur a cost to punish anti-social actions is one mechanism through which societies can promote pro-social behavior (Yamagishi, 1986; Henrich et al., 2006; Fehr and Gächter, 2002; Fehr and Fischbacher, 2004; Carpenter and Matthews, 2012; Charness et al., 2008). However, identifying the exact role of voluntary punishment in promoting cooperation and pro-sociality requires an improved understanding of the robustness of punishment behavior. Recent research on "moral wiggle room" and "reluctant altruism"—i.e., the tendency to voluntarily share wealth at a personal cost when confronted with a direct request, but to seek out excuses to avoid sharing altogether—shows that altruistic behavior, such as charitable giving, diminishes significantly if people can avoid the opportunity to give or possess excuses for not giving (Dana et al, 2006; Broberg et al., 2007; Dana et al., 2007; Haisley and Weber, 2010; Andreoni et al., 2011; Lazear et al., 2012; DellaVigna et al., 2012). These findings raise important questions for human social behavior more broadly *(*Malmendier et al., 2014).

In this paper, we investigate to what extent reluctance also applies to costly punishment. That is, when people are provided with opportunities to avoid engaging in costly punishment, without having to explicitly state that they do not want to punish, does the frequency of punishment of norm violation decrease? Moreover, does the answer to this question vary by whether the punisher is an independent outsider (third party) or a victim (second party) of the original transgression? The extent to which economic agents are reluctant punishers may provide insights into when we are likely to observe punishment in natural contexts and into the extent to which voluntary punishment can be an effective instrument for supporting pro-social and efficiency enhancing behavior.

To see why, consider how different motives may yield very different punishment behaviors and outcomes in the presence of opportunities to avoid carrying out punishment acts. On one hand, people may be *resolute* punishers of wrongdoing in the sense that they derive positive utility from punishing those who transgress norms, even when doing so is costly and it is possible to avoid the punishment opportunity. Hence, for some people, the satisfaction from punishing wrongdoing may exceed the costs of actually carrying out the punishment, thus making these people steadfast in seeing through punishment opportunities. For example, it may be that observing anti-social behavior arouses negative emotions such as anger, which may be

relieved through the act of punishing. Although behavioral and neuro-imaging studies show that people feel angry when being treated unfairly and that such negative emotion triggers second-party punishment (Fehr and Gachter, 2002; Sanfey et al., 2003; Xiao and Houser, 2005), the role of anger in third-party punishment remains unclear (Carlsmith et al., 2002; Batson et al., 2007; Pedersen et al. 2013; Jordan et al. 2015).

On the other hand, people facing the opportunity to punish wrongdoers may be *reluctant* punishers because they feel obligated to do so by social and ethical norms. A social obligation to punish may explain findings that publicity and anticipated guilt increase third party punishment behavior (Kurzban et al., 2007; Piazza and Bering, 2008) and that opportunities to compensate the victims (Bicchieri and Chavez, 2013) or to withhold help to the transgressors (Nikiforakis and Mitchell, 2014; Balafoutas et al. 2014) often reduce the frequency of punishment. The desire to be respected and evaluated positively by both others and even one's self can provide a unifying interpretation for complex aspects of human pro-social behavior (Akerlof and Kranton, 2000; Benabou and Tirole, 2006; Ellingsen and Johannesson, 2011; DellaVigna et al., 2012).[1] When punishing misconduct is socially desirable, an individual might feel compelled to incur costs to punish wrongdoers, without really wanting to do so. Hence, such an individual may also avoid carrying out costly punishment when an opportunity arises to do so while nevertheless stating a desire to punish. In view of the extensive literature demonstrating reluctance in other costly pro-social behaviors, costly punishment—especially third party punishment—might also be reluctant. Hence, such punishment may be sensitive to whether potential punishers can metaphorically "turn a blind eye"—that is, hide behind circumstances that allow them to avoid the punishment opportunity.

To test and compare the robustness of punishment behavior of both second and third parties, we conducted three variants of a dictator game experiment with punishment opportunities. Our primary purpose was to examine the extent to which third parties, as compared with second parties, choose to capitalize on an opportunity to secretly limit their ability to act on a stated intention to punish a wrongdoer. That is, to what extent do second- and third-party punishers demonstrate reluctance or resoluteness in their punishment behavior, when

[1] Van der Weele et al. (2014) show that the availability of an excuse has no effect on the reciprocal behavior in a context involving positive reciprocity—specifically a trust game—and argue that image concerns are not a key driver of reciprocal behavior. In contrast, Malmendier et al. (2014) find evidence of reluctance in reciprocal behavior that is similar in magnitude to the reluctance in one-sided dictator games.

confronted with an opportunity to avoid having to carry out a costly punishment act while "saving face" by nevertheless expressing an intention to punish?

Previous laboratory experiments studying voluntary punishment typically provide decision makers with clear choices (e.g., how much to punish) and direct mappings from choices into outcomes. Our experiment differs from these previous experiments in that we allow the decision maker to indicate the intention to punish but also subsequently to manipulate the *ability* to punish without doing so transparently. Specifically, punishers first decide whether and how much to punish the dictator, and such "intended" punishment decisions are visible to all relevant parties. But we then allow a random outcome to determine whether punishment decisions can actually be enacted, and let those making the punishment decisions secretly misreport the random outcome. By doing so, our experiment allows us to examine the extent to which punishing behavior is reluctant or resolute. In our experiment, a resolute punisher will actively seek the opportunity to punish, even if punishment is costly; a reluctant punisher, after demonstrating the intention to punish, will try to avoid the opportunity to implement costly punishment when excuses for doing so are available.

This design is a step closer to many natural settings in which individuals can often manipulate the circumstances and context (Dana et al, 2006; Dana et al., 2007; Lazear et al, 2012). As an example of resolute punishment, a disgruntled employee who feels personally wronged by co-workers or an employer may go out of his way to cause them harm—perhaps stalking them or seeking opportunities for a confrontation. An extreme instance of such behavior is the murder by Vester Lee Flanagan II in 2015 of two of his former co-workers at a local news station who he felt had wronged him. On the other hand, a reluctant punisher may pretend not to notice transgressions when it may be personally costly to act to punish them. This is what the Freeh Report, commissioned by the Board of Trustees of the Pennsylvania State University, concluded was done by high-ranking university officials, including President Graham Spanier and football coach Joe Paterno, in response to child sexual molestation by a prominent assistant coach, Jerry Sandusky.

Our experimental findings provide clear evidence for the reluctance of third-party punishment and for the resoluteness of second-party punishment. Many third parties who express a willingness to engage in costly punishment nevertheless exploit a loophole in the environment in order to avoid actually having to carry out the punishment act. However, second parties

exhibit the opposite pattern in their behavior—they resolutely seek out the ability to punish, going as far as to misreport random outcomes in order to ensure that the punishment act is carried out. That is, people tend to turn a blind eye to the wrongdoers when they are unaffected observers but do not turn the other cheek, accepting the transgression without any retaliation, when they are the victims.[2]

Our findings yield two important insights regarding the punishment behavior of economic agents. First, we provide clean empirical evidence that sheds light on the differences in the nature of the motives underlying second- and third-party punishment. Punishment by those who have been personally wronged may reflect more of a "true" or "internal" preference, while punishment on behalf of others appears to be driven more by external considerations, such as social pressure or image concerns. Second, our findings also raise important issues for the design of institutions in which second- and third-party punishment are possible, but where people may have the ability to avoid the opportunity or obligation to punish. The reluctance of punishment by third parties may make it infrequent and may diminish the extent to which it is an instrument that discourages anti-social behavior. Conversely, however, retaliation by those personally wronged may be widespread—and may even occur in situations where it is socially undesirable (Houser and Xiao, 2010).

2. Experimental Design

Our experiment is a revised version of prior experiments on second- and third-party punishment (Fehr and Fischbacher, 2004) and consists of two stages. In Stage 1 (the dictator game stage), one party has the option to voluntarily share or not share with a second party.[3] Stage 2 (the punishment stage) consists of two parts. In the first part, the opportunity to punish the first party—at a cost—is offered to either the directly affected second party or to a neutral third party, depending on the treatment condition. In the second part of the punishment stage, a random process determines whether the punishment decision in the first part is implemented. However,

[2] The idiom "turn a blind eye" means to pretend not to notice and is said to have originated when, during the Battle of Copenhagen (1801), Admiral Horatio Nelson held a telescope to his one blind eye and declared that he did not see the signal flags ordering him to "discontinue the action." The phrase "turn the other cheek" means to refrain from retaliation and alludes to the words of Jesus as described in the New Testament, Matthew 5:39 KJV.

[3] One advantage of using a dictator game to address our research question is that it is a fairly simple game. Given the complication of the punishment implementation stage in our experiment, it is important to keep the game as simple as possible.

our two main treatments provide subjects the opportunity to misreport the random outcome, thereby avoiding having to carry out the costly punishment.

2.1 Experimental Conditions

The experiment consists of two principal experimental conditions: *third-party punishment with self report* (3P Self Report) and *second-party punishment with self report* (2P Self Report). In addition, we conducted a control condition, *third-party punishment with verification* (3P Verified), to address a potential interpretation of behavior in the 3P Self Report condition.[4]

In the first condition, 3P Self Report, three parties form a group. Group members make decisions across two stages.

In Stage 1, the first party (Participant A) has the opportunity to act fairly or unfairly toward the second party (Participant B), in allocating 100 points between the two participants. In each group, Participant A decides whether to give 0, 10, 20, 30, 40, or 50 points to Participant B, keeping the remainder of the 100 points for him or herself. A third party (Participant C) receives a fixed amount of 50 points. Points are worth 10 cents each.

Stage 2 is the punishment stage and consists of two parts. In the first part, we elicit punishment strategies from the third party, Participant C. Specifically, for each of the possible actions taken by the first party in the first stage, Participant C chooses how many deduction points (between 0 and 50) to assign to Participant A. Participant C is also told that all his/her decision will be shown to Participants A and B at the end.[5] Following Fehr and Fischbacher (2004), each assigned deduction point reduces Participant C's final earnings by one point and Participant A's final earnings by three points. Participant B's earnings are unaffected by Participant C's punishment decision. Such designs have previously demonstrated a willingness to punish among significant proportions of third parties (Fehr and Fischbacher, 2004).

[4] Instructions for all three treatments are provided in Appendices A, B, and C.

[5] We attempted to use a design as similar as possible to Fehr & Fischbacher (2004), including using the strategy method to elicit punishment choices. The strategy method is desirable because it allows us to obtain meaningful observations on punishment behavior in cases where Participant A acts fairly. One main disadvantage of using strategy method is that it might diminish the role of emotion in decision making. If emotion plays a more important role in second-party punishment than in third-party punishment, the strategy method could lead to less punishment driven by emotion and make it harder to observe the difference that we find between the second- and the third-party punishment treatments. Thus, the differences we observe in our data might be viewed as a lower bound of the difference between second- and third-party punishment.

The novel feature of our design is the second part of Stage 2. This part mimics aspects of many natural environments, whereby third parties may vary in their actual opportunity to execute the intent to punish a wrongdoer, and may have some ability to manipulate this opportunity. For example, a transgressor may escape before a third party can act, or a third party may simply not observe or be aware of the transgression. Moreover, such contextual features may also allow third parties the possibility of misrepresenting whether they are actually limited in their ability to punish—or to "turn a blind eye" to a potentially punishable transgression.

We implement this feature in the second part of Stage 2 by having the third party roll a die to determine whether or not the punishment decision made in the first part of Stage 2 is actually enacted. In particular, each Participant C receives a cup with a die inside. The experimenter then announces publicly, "If you are Participant C, please roll the die privately in your cup at least five times to verify that it is a fair six-sided die. Once you have rolled it enough times to be satisfied it is a fair die, roll it privately one more time. You do not need to announce when you are rolling for the final time. Then report the result of your final roll below and click the 'Submit' button." Participants were told that an even number in the final roll corresponds to states of the world in which the punishment is actually carried out, while an odd number creates outcomes in which no punishment is enacted and the third party incurs no cost. Thus, if Participant C rolls and reports the results in an unbiased manner, then there is a 50% probability that any attempt to punish by Participant C is irrelevant and the outcome is identical to one in which there is no possibility of punishment.

Importantly, note that the design calls for Participant C to roll the die several times privately, to decide when to roll one final time, and to only report the outcome of the last die roll. That is, Participant C can decide privately and independently when to stop, and which die roll will "count." By deciding on their own which roll is final, participants can not only keep the truth secret but also potentially *manipulate the truth*. That is, our design allows participants to deceive not only others but also themselves. A reluctant punisher who does not truly desire to punish but does so only due to image concerns, *even self-image concerns,* can represent a desire to punish a transgressor (despite the cost) in the first part of Stage 2, while manipulating or misrepresenting the outcome of the die roll to produce an odd number, so the costly punishment does not actually take place and the punisher incurs no cost in the end. On the other hand, a resolute punisher who truly desires to punish the wrongdoer can specify the decision to incur the

cost to punish and then manipulate or misreport the die roll outcome to ensure that it is implemented.

By comparing the distribution of reported die rolls to the known distribution of outcomes of a fair die, we can identify whether the third parties are intentionally avoiding or seeking out the ability to enact costly punishment by selectively reporting the die roll outcomes that they prefer (Fischbacher and Heusi-Follmi, 2003; Shalvi et al., 2011). Of course, a preference for strictly acting honestly may still prevent an individual from manipulating or misrepresenting the die roll outcome. Therefore, for example, even if all third parties are reluctant punishers, we might observe less than 100 percent of punishing third parties reporting odd numbers simply because some reluctant punishers are also unwilling to manipulate the outcome for these stakes. Similarly, resolute punishers may not always report even numbers. Hence, we expect the proportion of realized odd die rolls to lie between 50 and 100 percent for reluctant punishers and between 0 and 50 percent for resolute punishers. For the purpose of our study, the key statistic for uncovering whether punishment is, on average, resolute or reluctant is whether the reported proportion of even rolls lies significantly above or below 50 percent.

Our other principal condition, second party with self-report (2P Self Report), is a modification of 3P Self Report that allows us to explore the extent to which second-party punishment is resolute or reluctant—i.e., whether it is similarly sensitive to the opportunity to avoid acting on a stated preference for punishment. In 2P Self Report, there is no third party (Participant C), and it is Participant B who has the opportunity to specify punishment for Participant A and then privately roll a die in Stage 2. Hence, Participant B serves the role of both the recipient of the Stage 1 decision and potential punisher in Stage 2. The punishment decision and die roll are otherwise identical to those in the 3P Self Report condition. As with 3P Self Report, the distribution of die rolls reported in 2P Self Report informs of us the nature of the punishment decisions. A reluctant second-party punisher may use the private die roll opportunity to report an odd-numbered outcome, and thereby avoid having to punish, while a resolute punisher determined to see the punishment through will report an even-numbered outcome.

Finally, we note that an alternative interpretation of a high proportion of reported odd-numbered die outcomes is that individuals who would normally not incur a cost to punish state otherwise in order to voice their (costless) disapproval of the first party's actions (Xiao and Houser, 2005). They can guarantee that signaling such disapproval is costless by reporting an

odd-numbered outcome for the die roll in Stage 2. Thus, under this interpretation, a low frequency of even numbers reported does not inform us of the true preferences of the people who would otherwise punish, but instead reflects the actions of people who would not punish if there was no opportunity to misreport the die roll and who now claim that they would punish simply because they can get out of it. A key prediction under this interpretation is that the proportion of subjects who indicate a willingness to punish should be significantly higher when they can subsequently misreport the outcome of the die roll than when they cannot.

We report below that there is a high proportion of reported odd rolls in the 3P Self Report condition, but not in the 2P Self Report condition. It is therefore important to address the above possible alternative explanation only for the 3P Self Report results. For this reason, we conducted an additional treatment, third party with verification (3P Verified).[6] The only difference between the 3P Self Report and 3P Verified conditions is that, in the latter, the experimenter visually observed the third party's die roll and entered it into the computer. Hence, there was no opportunity to manipulate the outcome of the die roll. All subjects knew this in advance of their decisions.

2.2 General Procedures

We conducted experimental sessions at the Pittsburgh Experimental Economics Laboratory (P.E.E.L.). We targeted approximately 60 subjects in the role of punishers in each condition, but the actual number of observations varied slightly based on variance in recruiting and in the number of participants attending sessions. Specifically, the experiment consisted of 10 sessions of the 3P Self Report condition with a total of 174 subjects (58 punishers), 9 sessions of the 3P Verified condition with a total of 165 subjects (55 punishers), and 7 sessions of the 2P Self Report condition with a total of 122 subjects (61 punishers).[7]

---

[6] We ran 3P Verified treatment after we obtained evidence of a high proportion of reported odd rolls (i.e. not to implement the stated punishment) in the 3P Self Report condition, but not in the 2P Self Report condition. Because the Verified condition is designed only to test the alternative explanation of the high proportion of reported odd rolls, we did not run a 2P Verified treatment.

[7] These numbers exclude one pair in the 2P Self Report condition in which the subject in the role of Participant B opted to leave the experiment after the instructions were completed but before any decisions were made. We targeted 60 subjects with the opportunity to punish in the 3P-Self-Report and 2P-Self-Report conditions. With 60 subjects, the empirical frequencies of punishment from Fehr and Fischbacher (2004) yield, in expectation, 37 third-party punishers (60*.61) and 44 second-party punishers (60*.74). In the former case, this gives us 94 percent power to detect a proportion that differs from 50 percent by +/- 25 percent at $p < 0.05$ in a two-sided test; in the latter case we have 97 percent power (Chow, Shao & Wang, 2008).

Each session lasted approximately 45 minutes. The experiment was conducted using the software z-Tree (Fischbacher, 2007). Following role assignment and instructions, subjects completed a quiz through the computer interface to verify understanding of the instructions before making any decisions. Once all subjects successfully completed the quiz, the computer monitor showed them the decision screens relevant to their role. Subjects in the role of dictator made their allocation decisions. Then, subjects in the role of punisher specified, for each possible amount shared by the dictator, how many deduction points to assign. Once all punishment strategies were elicited, the experimenter distributed to each potential punisher a die inside a cup and instructed them to use this die to determine whether the punishment decision would be implemented. In the 3P Verified condition, the experimenter went to each subject in the role of punisher sequentially, observed the die roll, and entered this outcome into the computer.

At the end of the experiment, subjects saw all the decisions made by each party, the reported die roll, and final outcome for their group, but received no information on what happened in other groups. Subjects were informed during the instructions that in the event that a player ended the experiment with negative points, the negative earnings would be deducted from the $8 payment each subject received for participating in the experiment.[8]

3. Results

Our primary interest is in the frequencies with which subjects with the possibility to punish state the intention to do so, and subsequently enact the stated punishment decision in the conditions with self-reported die rolls. Providing an overview of the results, Figure 1 presents the frequencies of stated punishment intentions and enacted punishment across the three treatment conditions.

3.1 Robustness of Third-party Punishment

Our first finding is that reluctance appears to be a central component of third-party punishment, as evidenced by the unusually low frequency of reported even die-roll outcomes. Of the 58 Participants C in the 3P Self Report condition, 23 (39.7 percent) made the initial decision, in

[8] Including the participation payment, mean earnings in the 2P Self Report condition were $14.25 for the first party and $9.70 for the second party. Means earnings in the 3P Self Report condition were $17.18 for the first party, $8.69 for the second party, and $12.96 for the third party. Mean earnings in the 3P Verified condition were $16.41 for the first party, $8.89 for the second party, and $12.77 for the third party.

Stage 2, to punish (at a cost to themselves) at least one of the potential actions of the first party.[9] If they were reporting die roll outcomes as the result of a natural and unbiased random process, we would expect 50% of these subjects to report an even number. However, only 5 of 23 third parties (21.7%) reported rolling an even number, which is statistically unlikely from an unbiased random process ($p < 0.02$ in a binomial test). Thus, consistent with reluctance in third party punishment, we find that third parties who expressed a willingness to incur costs to punish transgressing first parties also tended to misrepresent the outcome of the die roll in a manner that "prevented" them from being able to actually implement this costly action. As a result of such reluctance, and the manipulation opportunity afforded by the private die roll, third party punishment is rare in the 3P Self Report condition: only 5 of 58 third parties (8.6 percent) effectively implemented punishment of the first party.

In addition to reluctance, another possible interpretation of the high frequency of reported odd rolls is that some Participants C who would not otherwise choose to incur a cost to punish Participants A in Stage 2 can now do so costlessly—e.g., to express their disapproval—by simply reporting an odd die roll outcome in the second part of Stage 2. Data from the 3P Verified treatment do not support this alternative explanation. We find that 19 of 55 (34.5%) Participants C in this condition chose to punish at least one of the actions of the first party, which is very close to and does not significantly differ from the 39.7 percent who chose to punish in the 3P Self Report condition (fisher's exact test, $p=0.70$). The punishment rate of 34.5 percent in the 3P Verified treatment implies an expected effective punishment rate of 17.3 percent—one half of the total choosing to punish—which is twice as high as the effective punishment rate by third parties in the 3P Self Report condition (8.6 percent)[10].

Hence, the low rate of reported even rolls in 3P Self Report does not appear to be driven by additional Participants C indicating they would punish only because they knew they could avoid it subsequently. Most of the difference in behavior between these two conditions lies in the implementation of the punishment, rather than in the statement of an intention to punish. This is further supported by regression analyses, discussed in Section 3.3, which show that the verified die roll has no significant effect on third parties' Stage 2 probability of stating a preference for

[9] This frequency of costly third party punishment is lower than has been found in some prior experiments. For example, Fehr and Fischbacher (2004) found 61% of third parties (and 74% of second parties) were willing to punish unfair actions.

[10] As expected, the realization rate of intended punishment in 3P Self Report condition does not statistically differ from 50% ($p=0.17$) in a binomial test.

punishment, the average amount of punishment among those that do punish, or their sensitivity to the amount Participant A gives. That is, the stated punishment intention is very similar between the 3P Self Report and 3P Verified conditions. Therefore, we attribute the approximate halving of enacted punishment to the fact that Participants C had the ability to act on their true preferences. That is, the veil of randomness provided by the die roll shields reluctant third-party punishers from social image concerns and allows them to "turn a blind eye" to the first party's transgression, thus circumventing the responsibility to punish.

3.2 Robustness of Second-party Punishment

The 2P Self Report condition allows us to explore the extent to which second-party punishment is resolute or reluctant—i.e., whether it is similarly sensitive to the opportunity to avoid acting on a stated preference for punishment, as is the case with third-party punishment. As shown in Figure 1, 35 of 61 (57.4 percent) second parties chose to punish at least one of the possible actions by the first party. Of these 35 subjects who elected to punish, 24 (68.6 percent) reported rolling an even number and therefore had their punishment decision enacted. This proportion is statistically significantly greater than the 50 percent we would expect if second parties were honestly reporting the outcome of the die roll ($p < 0.05$ in a binomial test). Thus, in contrast with the reluctance exhibited by third parties, second-party punishers are not willing to "turn the other cheek" and refrain from retaliation after being treated unfairly. Instead, not only do they not use the die roll as an excuse to avoid implementing the punishment, some of the second parties even appear to be willing to misreport the die roll to ensure the punishment is enacted.[11]

3.3 Punishment Amounts

In addition to the decision of whether to punish, we also examine the punishment amount assigned by second- or third-party punishers. Figure 2 plots the mean stated punishment expenditures, for those subjects who indicated a willingness to punish at least one of the possible decisions by Participant A. There is a clear difference in the punishment intensity between

[11] The preceding analyses focus on the frequency of subjects who chose to punish at least one of the possible actions by the first party, which gives us the largest number of stated punishment decisions. To check the robustness of the results, we also examined the frequency of those who punish at least twice, three times, and so on. We obtain the same qualitative results—proportions of enacted punishment vary between 15 and 27 percent for third parties and between 69 and 72 percent for second parties. However, the samples are smaller and the significance levels of the binomial tests are sometimes weaker. All the results are reported in Appendix D.

second and third parties. Specifically, for all actions by Participant A that depart from fairness, second parties who punish do so more intensely, on average, than do third parties. Overall, the mean of implemented punishment amounts is significantly lower in the 3P Self Report than that in the 2P Self Report treatment (2.45 vs. 7.48, $t_{117}=3.445$, $p<0.001$).[12]

Importantly, when we look at punishment amounts in the two versions of the experiment with third-party punishment, we do not observe a difference between the punishment behavior of those third parties whose die rolls would be verified (3P Verified) and those who were free to report any outcome (3P Self Report). There is no significant difference in the average punishment amount between the two treatments (2.45 vs. 2.08, $t_{111}=0.414$, $p=0.68$). This supports our earlier conclusion that expressed third party intentions do not differ between the two conditions.

Further evidence that the ability to subsequently reverse stated punishment intentions did not influence punishment behavior is observed when comparing those who ultimately reported the ability to punish (even-numbered die rolls) with those who ultimately reported an inability to punish (odd-numbered die rolls). Figure 2 reveals no differences in intended intensity of punishment for the two groups (3P Self Report: Mean punishment of 5.13 for even die rolls vs. 6.46 for odd die rolls, Mann-Whitney rank-sum test, $p=.79$; 2P Self Report: Mean punishment of 12.91 for even die rolls vs. 13.35 for odd die rolls, Mann-Whitney rank-sum test, $p=0.90$). For the 3P Self Report and 2P Self Report conditions, the lack of a difference is meaningful—it indicates that the decision of whether and how much to punish, made in the first part of Stage 2, is not influenced by whether one is likely to ultimately avoid implementing the punishment.

3.3. Modeling the Punishment and Die Roll Decisions

We next model the punishment and die-roll-reporting decisions, using regression analysis, as a way of understanding the statistical relationships observed in Figures 1 and 2. Table 1 compares the behavior of punishers in the two conditions with self-reported die roll outcomes (3P Self Report and 2P Self Report). Models 1 and 2 address the decision of whether and how much to punish in the first part of Stage 2 (i.e. the intended punishment). Then, in models 3 and 4, we

[12] We also observe some subjects (21 out of 174) that state a preference for punishment even when Participant A gives 50. We interpret this as the anti-social punishment discussed in previous literature (e.g. Hermann, Thöni, Gächter, 2008). We see 10 (of 61) such subjects in 2P Self Report, 3 (of 58) in 3P Self Report, and 8 (of 55) in 3P Verified.

analyze the subsequent second part decision regarding reporting of the die roll (i.e. the actual punishment).

Because the decision of *whether* to punish a wrongdoer is logically separable from the choice of the *amount* of punishment, we model the Stage 2 punishment decisions of our subjects in two stages. Such a hurdle model (Erkal et al., 2011) first examines the binary outcome of punishing or not punishing and then, if punishment is chosen, independently estimates the determinants of the amount of punishment selected. Because each subject makes six punishment decisions, one for each possible choice by Participant A, but observations are otherwise independent across subjects in the role of punisher, we include subject random effects.

In the first stage of the hurdle model (Table 1, model 1), we estimate a probit regression of the stated intention to punish on a binary variable identifying the 2P Self Report treatment, the amount that Participant A gives to Participant B (0, 10, 20, 30, 40, or 50), and the interactions between these. The dependent variable is the binary decision of whether to punish (1) or not (0). The 3P Self Report condition is the baseline against which the coefficient for 2nd Party Self Report measures differences. The results reveal that second parties are significantly more likely to state a preference for punishment when Participant A gives zero. As expected, we also observe that as the amount given by Participant A increases, the probability of punishing decreases. The sensitivity to giving by Participant A does not significantly differ between the 3P Self Report and 2P Self Report conditions, indicated by the statistically insignificant interaction term.

The second stage of the hurdle model (Table 1, model 2) consists of a truncated linear regression of the punishment amount on the same independent variables. Here, we observe that, relative to third parties, second parties assign more deduction points to Participant A. Directionally, more punishment points are assigned to those Participants A who shared less, but this is not statistically significant. We again observe that punishment is not significantly more sensitive to initial sharing for second parties than for third parties.

Models 3 and 4 study punishment enactment in the second part of Stage 2 (i.e., self-reported die rolls) for the 58 subjects who chose to punish in the first part (23 subjects in 3P Self Report and 35 in 2P Self Report). We confirm that second parties who chose to punish are significantly more likely to report die rolls that lead to punishment being enacted. In neither self-report treatment is the mean punishment expenditure significantly predictive of whether punishment is enacted. That is, consistent with Figure 2, those people who subsequently enacted

punishment did not systematically indicate higher or lower punishment amounts in the first part of Stage 2.

The conclusion of this analysis is that there are significant differences in whether and how second and third parties enact costly punishment. Second parties punish both more frequently and more intensively.[13] More importantly, for our purposes, they are more resolute in their punishment—they report many more die roll outcomes that allow the punishment decision to be enacted.[14]

We also compare behavior in the two third-party punishment treatments, which vary based on whether the third party enacting the punishment can misreport the die roll (3P Self Report) or can not do so (3P Verified). Table 2 presents the hurdle model of the decision, in the first part of Stage 2, of whether or not to punish and, conditionally on doing so, of how much to punish. Consistent with our earlier analysis, we observe no significant difference between the two versions of the experiment with third-party punishers. While the probability of a third party punishing is sensitive to the amount given by Participant A, this sensitivity does not significantly depend on whether they will later be able to manipulate the outcome of the die roll (model 1). Similarly, the amount of punishment chosen by third parties does not depend on whether the die roll will be verified (model 2). Overall, the results in Table 2 provide strong support for our earlier observation that the decision of whether to *state* an intention to punish is not affected by whether the third party will subsequently have the opportunity to avoid implementing this decision by misreporting the die roll outcome.

4. Discussion

We provide evidence that the decisions of third parties to punish wrongdoers are often reluctant. That is, third parties who state an intention to punish norm transgressors often capitalize on the ability to avoid actually implementing the costly punishment act. In contrast, we find that second-party punishment is resolute. Second parties are not only more likely to pay a cost to

[13] First parties (dictators) seem to anticipate this difference. While their behavior does not significantly differ between the 3P Self Report and 3P Verified treatments (χ^2, p=.71), in the 2P Self Report treatment, dictators act more generously (χ^2, p<.01). While 64% of dictators give zero in the third party treatments, only 26% do so in 2P Self Report. The mean amount given in 3P Self Report and 3P Verified are 6.9 and 8.9, respectively, while the mean amount given in 2P Self Report is 22.1 (p<.001 for comparison to each 3P treatment).

[14] Note that the comparisons between the second party and the third party in the regression results reported allow three possibilities: misreport in 2P Self Report but not in 3P Self Report, misreport in 3P Self Report but not in 2P Self Report, and misreport in both. The binomial tests we reported in the above two sections indicate that both parties tend to misreport—i.e., the proportions for each group differ from 50 percent—but in the opposite directions.

punish the wrongdoer, but they also seek out the opportunity to implement punishment and this desire is sufficiently strong to induce some subjects to manipulate otherwise random events in order to ensure that punishment is implemented.

Our work highlights the need to better understand the deeper motives underlying social behavior (Malmendier et al., 2014; Kurzban et al., 2007). In the case of resolute versus reluctant punishment, one possibility is that the primary motivators for third-party (often reluctant) punishment tend to be extrinsic, higher-order, socially constructed concepts such as image or duty, which can be actively managed and manipulated in self-interested ways. Resolute second-party punishment, on the other hand, may be more intrinsically motivated, perhaps driven by emotions like anger.

Our findings also illuminate recent discussions on the difference between punishment observed in the lab and in the field (Guala, 2012; Balafoutas and Nikiforakis, 2012). Cooperation in a society that relies heavily on spontaneous third-party punishment may not be sustainable if situations arise in which individuals responsible for implementing punishment, and bearing its costs, have the ability to escape this social burden in a manner that also allows them to save face. The comparisons between 3P Self Report and 3P Verified conditions further suggest that exogenously enforcing the implementation of punishment intentions will not change the initial stated severity of punishment for wrongdoings. An implication of these findings is that it may be beneficial for a society to eliminate opportunities that excuse otherwise uninvolved third parties from the responsibility of defending social order. In some cases, this may require additional incentives, such as rewards or legal obligations, to compel people to take costly actions to sanction norm violations when they are not directly impacted by a transgression. Hence, "good Samaritan" or "duty to rescue" laws, which diminish the reasons one can provide as exculpation for not having intervened in rendering aid, may partly reflect reactions to justifications employed by reluctant third parties confronted with opportunities to take costly acts that assist others.

On the other hand, the greater resoluteness of second-party punishment reveals that those directly wronged by a transgression are willing to bend moral rules—in our case, dishonestly reporting the outcome of a die roll—to ensure the enactment of punishment. Hence, this suggests that institutions should be designed to ensure that, for instance, such a resoluteness does not compromise efficiency (Dreber et al., 2008; Rockenbach and Milinski, 2006; Egas and Riedl, 2008; Xiao and Kunreuther, forthcoming) and lead to escalation of conflicts.

More broadly, our results highlight the importance of better understanding the motives that underlie different forms of pro-social behavior, in order to more accurately understand the impacts they are likely to have on society. While the willingness to incur costs to punish norm violators appears to be a robust phenomenon in settings where such a preference is directly elicited, our findings indicate that studying the extent to which such a stated preference actually ends up impacting social outcomes requires using more complex choices that shed light on the reluctance or resoluteness underlying such apparent preferences.

References:

Akerlof GA, Kranton RE (2000) Economics and Identity. *Quarterly Journal of Economics* 115(3):715-753.

Andreoni J, Rao JM, Trachtman H (2011). Avoiding the Ask: A Field Experiment on Altruism, Empathy and Charitable Giving. NBER Working Paper 17648.

Balafoutas L, Nikiforakis N (2012) Norm Enforcement in the City: A Natural Field Experiment. *European Economic Review* 56(8):1773-1785.

Balafoutas Loukas, Nikos Nikiforakis and Bettina Rockenbach. 2014. "Direct and indirect punishment among strangers in the field." *Proceedings of the National Academy of Sciences* 111 (45), 15924-15927.

Batson CD, Kennedy CL, Nord L, Stocks, EL, Fleming DA, Marzette CM, Lishner DA, Hayes RE, Kolchinsky LM, Zerger T (2007) Anger at unfairness: is it moral outrage? *European Journal of Social Psychology* 37(6):1272-1285.

Benabou R, Tirole J (2006) Incentives and Prosocial Behavior. *American Economic Review 95*(6):1652–1678.

Bicchieri, C. Chavez, A. (2013) Third-Party sanctioning and compensation behavior: Findings from the ultimatum game" Journal of Economic Psychology 39: 268-277

Broberg T, Ellingsen TM, Johannesson M (2007) Is Generosity Involuntary? *Economics Letters* 94(1):32-37.

Carlsmith K, Darley J, Robinson P (2002) Why do we punish? Deterrence and just deserts as motives for punishment. *Journal of Personality and Social Psychology* 83(2): 284-299.

Carpenter JP, Matthews PH (2012) Norm Enforcement: Anger, Indignation or Reciprocity? *Journal of the European Economic Association* 10(3):555-572.

Charness G, Cobo-Reyes R, Jimenez N (2008) An investment game with third-party intervention. *Journal of Economic Behavior & Organization* 68(1):18-28.

Chow S, Shao J, Wang H. 2008. *Sample Size Calculations in Clinical Research. 2nd Ed.* Chapman & Hall.

Dana J, Cain DM, Dawes RM (2006) What you don't know won't hurt me: Costly (but quiet) exit in dictator games. *Organizational Behavior and Human Decision Processes 100*(2):193–201.

Dana J, Weber RA, Kuang J (2007) Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness. *Economic Theory 33*(1):67–80.

DellaVigna S, List J, Malmendier U (2012) Testing for Altruism and Social Pressure in Charitable Giving, *Quarterly Journal of Economics* 127:1-56.

Dreber A, Rand DG, Fudenberg D, Nowak MA (2008) Winners Don't Punish. *Nature* 452:348-351.

Egas M, Riedl A (2008) The economics of altruistic punishment and the maintenance of cooperation. *Proceedings of the Royal Society B – Biological Sciences* 275(1637):871-878.

Ellingsen T, Johannesson M (2011) Conspicuous Generosity. *Journal of Public Economics* 95(9-10):1131-1143.

Erkal N, Gangadharan L, Nikiforakis N (2011) Relative Earnings and Giving in a Real-Effort Experiment. *American Economic Review* 101(7):3330-48.

Fehr E, Gachter S (2002) Altruistic punishment in humans. *Nature* 415(6868): 137-140.

Fehr E, Fischbacher U (2004) Third-party punishment and social norms. *Evolution and Human Behavior* 25(2):63-87.

Fischbacher U (2007) z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics* 10(2):171-178.

Fischbacher U, Heusi-Follmi F (2013) Lies in disguise: An experimental study on cheating. *Journal of the European Economic Association* 11(3):525-547.

Guala F (2012) Reciprocity: Weak or Strong? What Punishment Experiments Do (and Do Not) Demonstrate. *Behavioral and Brain Sciences* 35:1-15.

Henrich J, et al. (2006) Costly punishment across human societies. *Science* 312(5781):1767-1770.

Herrmann, Benedikt, Christian Thöni, and Simon Gächter. "Antisocial punishment across societies." *Science* 319.5868 (2008): 1362-1367.

Haisley E, Weber RA (2010) Self-serving interpretations of ambiguity in other-regarding behavior. *Games and Economic Behavior* 68(2):634-645.

Houser D, Xiao E **(**2010) "Inequality seeking punishment", *Economics Letters,* 109(1):20-23.

Jordan, J. J. and McAuliffe, K. and Rand, D.G. (2005) The Effects of Endowment Size and Strategy Method on Third-Party Punishment. Working paper.

Kurzban R, DeScioli P, O'Brien E (2007) Audience effects on moralistic punishment. *Evolution and Human Behavior 28*(215):75–84.

Lazear EP, Malmendier UM, Weber RA (2012) Sorting in experiments with application to social preferences. *American Economic Journal: Applied Economics 4*(1):136–163.

Malmendier UM, te Velde VL, Weber RA (2014) Rethinking reciprocity. *Annual Review of Economics*, 6:849-874

Nikiforakis Nikos and Helen Mitchell. 2014. "Mixing the carrots with the sticks: Third party punishment and reward" *Experimental Economics* 17 (1), 1-23.

Pedersen, E. J., Kurzban, R., McCullough, M. E. (2013). Do humans really punish altruistically? A closer look. *Proceedings of the Royal Society B: Biological Sciences*, 280, 20122723. http://dx.doi.org/10.1098/rspb.2012.2723.

Piazza J, Bering JM (2008) The Effects of Perceived Anonymity on Altruistic Punishment. *Evolutionary Psychology* 6(3):487-501.

Rockenbach B, Milinski M (2006) The efficient interaction of indirect reciprocity and costly punishment. *Nature* 444:718-72.

Sanfey AG, Rilling JK, Aronson JA, Nystrom LE, Cohen JD (2003) The Neural Basis of Economic Decision-Making in the Ultimatum Game. *Science* 300:1755-1758.

Shalvi S, Dana J, Handgraaf MJJ, De Dreu CKW (2011) Justified Ethicality: Observing Desired Counterfactuals Modifies Ethical Perceptions and Behavior. *Organizational Behavior and Human Decision Processes* 115(2): 181-190.

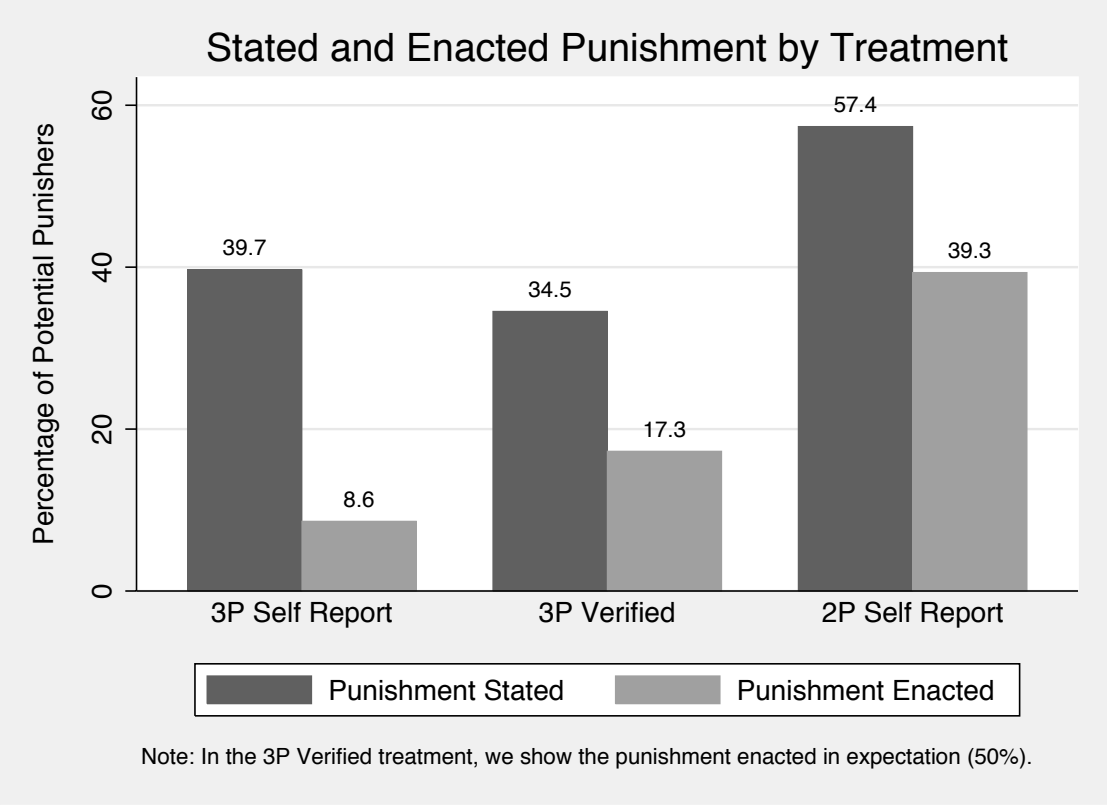
van der Weele JJ, Kulisa J, Kosfeld M, Friebel G (2014) Resisting Moral Wiggle Room: How Robust Is Reciprocal Behavior? *American Economic Journal: Microeconomics* 6(3): 256-64.

Xiao E, Houser D (2005) Emotion Expression in Human Punishment Behavior. *Proceedings of the National Academy of Science* 102(20):7398-7401.

Xiao E, Kunreuther H. (forthcoming) Punishment and Cooperation in Stochastic Prisoner's Dilemma Game. *Journal of Conflict Resolution.*

Yamagishi T (1986) The provision of a Sanctioning System as a Public Good. *Journal of Personality and Social Psychology* 51(1):110-116.

Fig. 1. The stated and enacted punishment by potential punishers in each of the three treatments.



Note: In the 3P Verified treatment, we show the punishment enacted in expectation (50%).

Fig. 2. Stated punishment expenditures by third-party subjects (Participant C) who expressed a preference for punishing at least one possible action by Participant A.

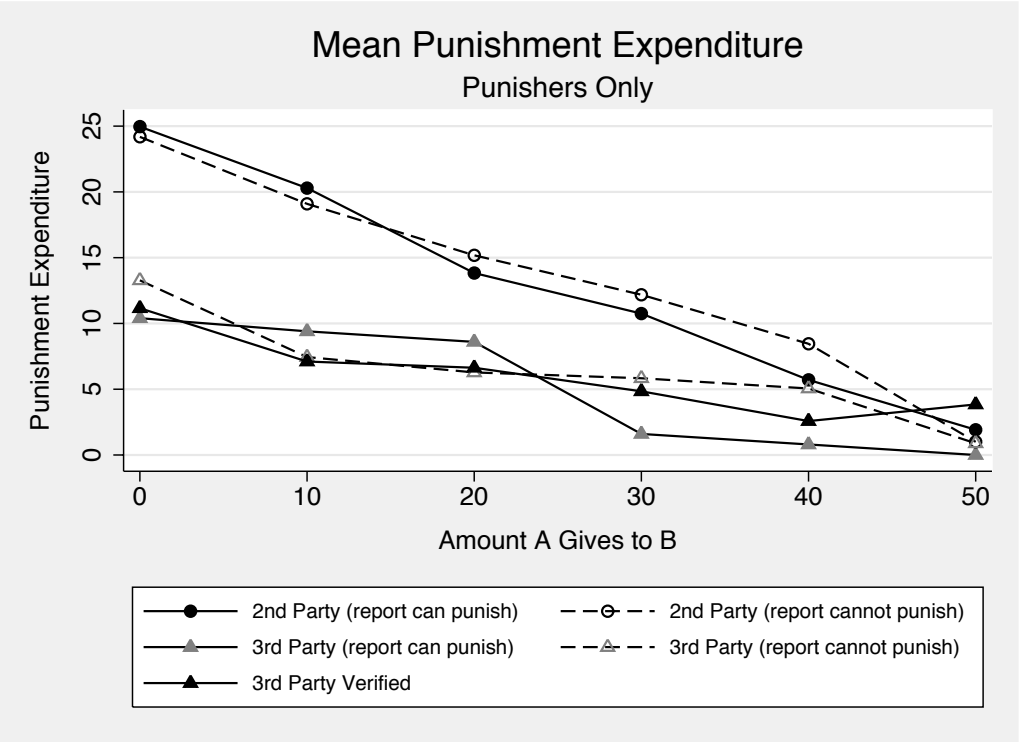


Table 1: Punishment Decisions and Die Rolls in Self-Report Treatments

	Stage 2 decisions			
	First part (stated intention to punish)		Second part (actual punishment)	
	(1) Probability of punishing	(2) Magnitude of punishment	(3) Enact Punishment	(4) Enact Punishment
2ndParty – Self Report	1.118**	17.74**	1.265***	1.157**
	(0.5059)	(7.025)	(0.367)	(0.555)
Amount given by Participant A	-0.037***	-0.367		
	(0.0078)	(0.273)		
Amount given by Participant A X 2nd Party Self Report	0.0034	-0.272		
	(0.0098)	(0.343)		
Mean Punishment Expenditure				-0.024
				(0.054)
Mean Punishment Expenditure X 2nd Party Self Report				0.022
				(0.058)
Constant	1.1551***	8.972	-0.781	-0.639
	(0.382)	(5.481)	(0.292)	(0.422)
Observations	714	223	58	58
Log likelihood	-272.92	-741.87	-33.83	-33.72
χ^2	51.61	102.66	12.75	12.97

Models include 3P Self Report and 2P Self Report conditions only; models 1and 2 include subject random effects; models 3 and 4 are probit regressions of the probability of reporting an even die roll. *** p<0.01, ** p<0.05, * p<0.1

Table 2: Stated Intention of Punishment in Third Party Treatments

	(1) Probability of punishing	(2) Magnitude of punishment
3rdParty – Verified	-0.4959	-0.198
	(0.5545)	(6.559)
Amount given by Participant A	-0.0389***	-0.302
	(0.0081)	(0.218)
Amount given	0.0162	0.0585
X 3rd Party Verified	(0.0110)	(0.291)
Constant	-1.3637***	11.25**
	(0.416)	(4.402)
Observations	678	147
Log likelihood	-214.40	-445.20
χ^2	30.92	28.89

Model 1 reports a probit regression of the binary decision of whether to punish and model 2 reports the results of a truncated linear regression of punishment amount; both models include 3P Self Report and 3P Verified conditions only and include subject random effects: *** p<0.01, ** p<0.05, * p<0.1

Appendix A: Instructions (3rd Party Self Report)

General Instructions

This is an experiment in decision-making.

In addition to a $5 show-up payment and $3 participation payment, you will be paid any money you accumulate from the experiment that will be described to you in a moment. You will be paid privately, in cash, at the conclusion of the experiment. The exact amount you receive will be determined during the experiment and will depend on your decisions and the decisions of others. *If you have any questions during the experiment, please raise your hand and wait for an experimenter to come to you. Please do not talk, exclaim, or try to communicate with other participants during the experiment.* Participants intentionally violating these rules will be asked to leave the experiment and will not be paid their earnings from the experiment.

During the experiment, payoffs will be expressed in points, not dollars. At the end of the experiment, points will be converted to dollars at the rate of $1 for every 10 points.

In this experiment, there are three types of participants: Participants A, Participants B, and Participants C. The computer will randomly and anonymously match participants into groups of three with one participant of each type. You will never be informed of the identity of the two people with whom you are matched, either during or after the experiment. Similarly, the two other participants in your group will never be informed of your identity. You will be in the same three-person group for the entire experiment.

Today's experiment consists of two stages. Any money that you earn from the experiment will be added to your $5 show-up payment and $3 participation payment. At the conclusion of the second stage, the experiment will end and you will receive your earnings in cash.

Stage 1 Instructions

At the beginning of this stage, the participants will receive the following number of points:

- Participant A: 100 points
- Participant B: 0 points
- Participant C: 50 points

In this stage, only Participant A will make a decision. Participant A will decide how many of the 100 points to give to Participant B. Participant A can give Participant B between 0 and 50 points in increments of 10 points. That is, Participant A can choose to give 0, 10, 20, 30, 40, or 50 points to Participant B.

> For example, if Participant A gives 40 points to Participant B, Participant A will have 60 points (100 – 40) and Participant B will have 40 points (0 + 40) at the end of this stage. If Participant A gives 10 points to Participant B, Participant A will have 90 points (100 – 10) and Participant B will have 10 points (0 + 10) at the end of this stage. If Participant A gives 0 points, Participant A will have 100 points and Participant B will have 0 points at the end of this stage.

Stage 2 Instructions

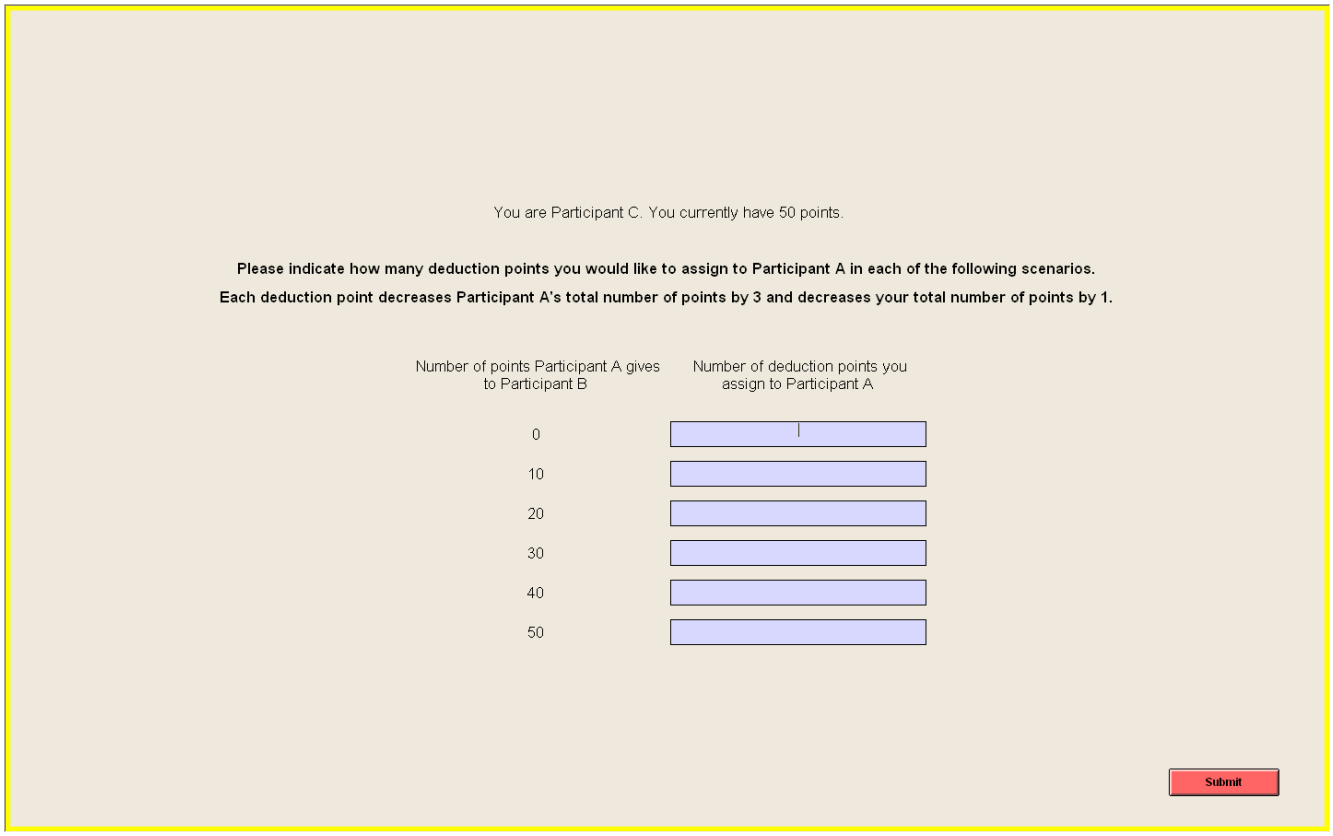
In this stage, only Participant C will make decisions.

Assigning deduction points

For each possible amount of points that Participant A gave to Participant B in Stage 1, Participant C will decide how many deduction points he or she would like to assign to Participant A. Participant C will make these decisions before knowing Participant A's actual choice in Stage 1. Participant C can assign between 0 and 50 deduction points to Participant A. For each deduction point that Participant C assigns to Participant A, Participant A's total number of points will be reduced by three and Participant C's total number of points will be reduced by one.

> For example, if Participant C assigns 2 deduction points, Participant A's total points will be reduced by 6 points (2x3) and Participant C's total points will be reduced by 2 points. If Participant C assigns 19 deduction points, Participant A's total points will be reduced by 57 points (19x3) and Participant C's total points will be reduced by 19 points.

Participant C will specify how many deduction points to assign to Participant A, for every possible amount that Participant A could give to Participant B in Stage 1. These decisions will be made on a screen like the one shown below.



Implementing the assignment of deduction points

After Participant C specifies the number of deduction points he or she would like to assign to Participant A for each possible decision by Participant A, a die roll will be used to determine whether or not Participant C's decision is actually implemented, and whether any deduction points are actually assigned to Participant A.

Participant C's decision will either

- *be implemented*, in which case the deduction points specified by Participant C, for the amount given by Participant A to Participant B, will be deducted from Participant C's total number of points and three times as many points will be deducted from Participant A's total number of points, or

- *not be implemented*, in which case no points will be deducted from either Participant C or Participant A.

Whether the decision is implemented will depend on a die roll reported by Participant C. Participant C will receive a paper cup containing a single six-sided die. Participant C will privately roll the die inside the cup, meaning that only Participant C will observe the outcome of the die roll. Participant C will then report the outcome of the die roll into the computer.

- If the reported die roll is an even number, meaning it is either **2, 4 or 6**, Participant C's decision about how many deduction points to assign to Participant A **will be implemented**. That is, Participant C's 50 points will be reduced by the specified number of deduction points and Participant A's total points will be reduced by three times this amount.

- If the reported die roll is an odd number, meaning it is either **1, 3 or 5**, Participant C's decision about how many deduction points to assign to Participant A **will not be implemented**. That is, Participant C will keep his or her entire 50 points and Participant A's total number of points will not be changed.

Points Calculation

The final number of points for each participant will be calculated as follows:

If Participant C's reported die roll is a 2, 4, or 6:

- Participant A's final number of points =
  + 100 (starting points)
  - Points Participant A gave to Participant B in Stage 1
  - 3 x Number of deduction points assigned to Participant A by Participant C in Stage 2 for the corresponding decision made by Participant A.

- Participant B's final number of points =
  + Points received from Participant A in Stage 1

- Participant C's final number of points =
  + 50 (starting points)
  - Number of deduction points assigned to Participant A in Stage 2 for the corresponding decision made by Participant A

If Participant C's reported die roll is a 1, 3, or 5:

- Participant A's final number of points =
  + 100 (starting points)
  - Points Participant A gave to Participant B in Stage 1

- Participant B's final number of points =
  + Points received from Participant A in Stage 1

- Participant C's final number of points =
  + 50 (starting points)

Your final number of points will be converted to dollars at the rate of $1 for every 10 points and added to your $5 show-up payment and $3 participation payment.

Please note that it is possible for Participant A's total number of points to be negative. In this case, the points will be deducted from Participant A's participation payment. Additional deduction points that would result in net losses greater than the $3 participation payment for Participant A will not count against the earnings of either Participant A nor Participant C.

Final Results

All decisions will be made through the computer. After all Stages 1 and 2 have finished, a results screen will reveal the following information:

- Participant A's decision of how many points to give to Participant B in Stage 1.
- Participant C's choice of deduction points in Stage 2.
- Participant C's reported die roll in Stage 2.
- Your final number of points.
- Your total earnings for the experiment.

Are there any questions about the instructions? If you have a question, please raise your hand and wait for the experimenter.

To make sure that everyone understands the instructions, we will now proceed to some questions about the instructions, which you will answer on the computer. You may refer to these printed instructions if you need to in order to answer the questions.

Appendix B: Instructions (2nd Party Self Report)

General Instructions

This is an experiment in decision-making.

In addition to a $5 show-up payment and $3 participation payment, you will be paid any money you accumulate from the experiment that will be described to you in a moment. You will be paid privately, in cash, at the conclusion of the experiment. The exact amount you receive will be determined during the experiment and will depend on your decisions and the decisions of others. *If you have any questions during the experiment, please raise your hand and wait for an experimenter to come to you. Please do not talk, exclaim, or try to communicate with other participants during the experiment.* Participants intentionally violating these rules will be asked to leave the experiment and will not be paid their earnings from the experiment.

During the experiment, payoffs will be expressed in points, not dollars. At the end of the experiment, points will be converted to dollars at the rate of $1 for every 10 points.

In this experiment, there are two types of participants: Participants A and Participants B. The computer will randomly and anonymously match participants into groups of two with one participant of each type. You will never be informed of the identity of the person with whom you are matched, either during or after the experiment. Similarly, the other participant in your group will never be informed of your identity. You will be in the same two-person group for the entire experiment.

Today's experiment consists of two stages. Any money that you earn from the experiment will be added to your $5 show-up payment and $3 participation payment. At the conclusion of the second stage, the experiment will end and you will receive your earnings in cash.

Stage 1 Instructions

At the beginning of this stage, the participants will receive the following number of points:

- Participant A: 100 points
- Participant B: 0 points

In this stage, only Participant A will make a decision. Participant A will decide how many of the 100 points to give to Participant B. Participant A can give Participant B between 0 and 50 points in increments of 10 points. That is, Participant A can choose to give 0, 10, 20, 30, 40, or 50 points to Participant B.

> For example, if Participant A gives 40 points to Participant B, Participant A will have 60 points (100 – 40) and Participant B will have 40 points (0 + 40) at the end of this stage. If Participant A gives 10 points to Participant B, Participant A will have 90 points (100 – 10) and Participant B will have 10 points (0 + 10) at the end of this stage. If Participant A gives 0 points, Participant A will have 100 points and Participant B will have 0 points at the end of this stage.

Stage 2 Instructions

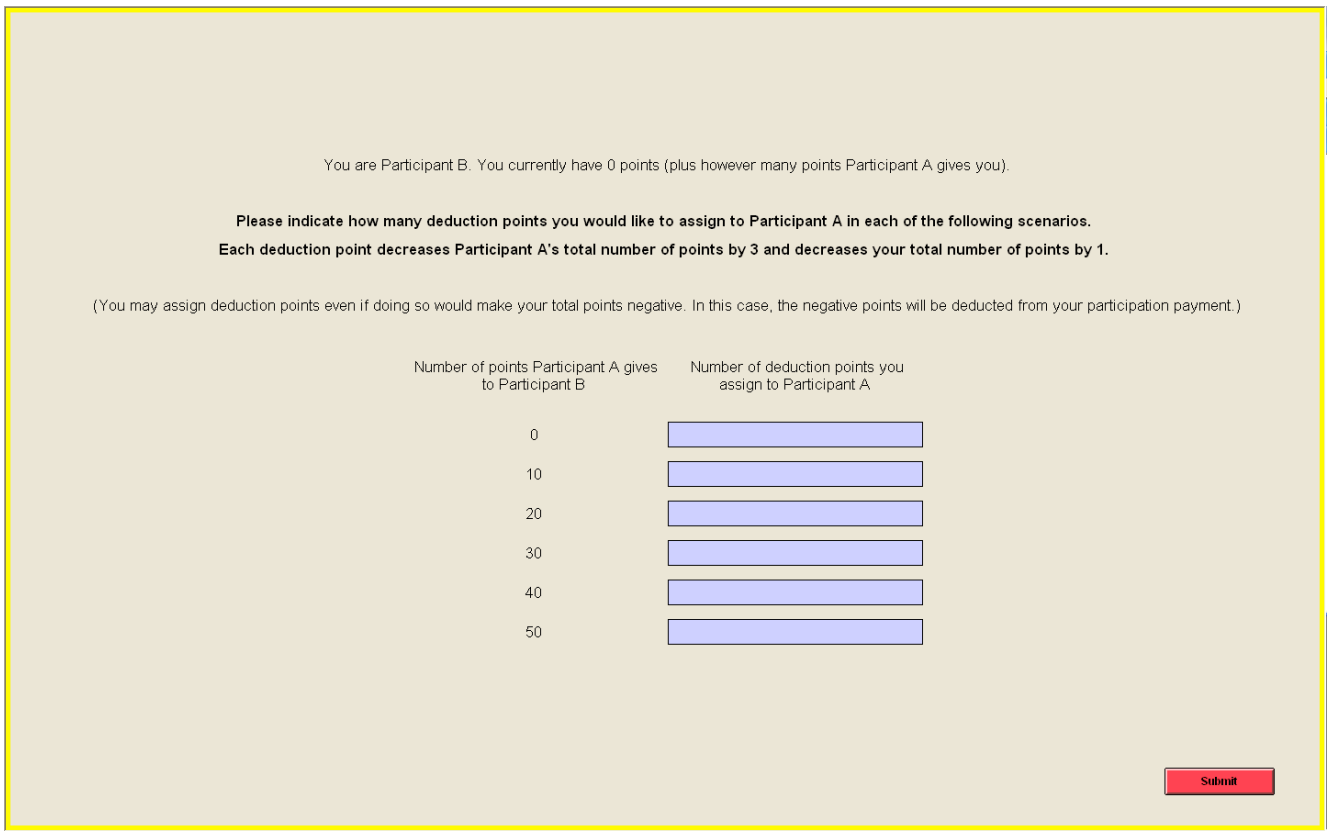
In this stage, only Participant B will make decisions.

Assigning deduction points

For each possible amount of points that Participant A gave to Participant B in Stage 1, Participant B will decide how many deduction points he or she would like to assign to Participant A. Participant B will make these decisions before knowing Participant A's actual choice in Stage 1. Participant B can assign between 0 and 50 deduction points to Participant A. For each deduction point that Participant B assigns to Participant A, Participant A's total number of points will be reduced by three and Participant B's total number of points will be reduced by one.

> For example, if Participant B assigns 2 deduction points, Participant A's total points will be reduced by 6 points (2x3) and Participant B's total points will be reduced by 2 points. If Participant B assigns 19 deduction points, Participant A's total points will be reduced by 57 points (19x3) and Participant B's total points will be reduced by 19 points.

Participant B will specify how many deduction points to assign to Participant A, for every possible amount that Participant A could give to Participant B in Stage 1. These decisions will be made on a screen like the one shown below.



Implementing the assignment of deduction points

After Participant B specifies the number of deduction points he or she would like to assign to Participant A for each possible decision by Participant A, a die roll will be used to determine whether or not Participant B's decision is actually implemented, and whether any deduction points are actually assigned to Participant A.

Participant B's decision will either

- *be implemented*, in which case the deduction points specified by Participant B, for the amount given by Participant A to Participant B, will be deducted from Participant B's total number of points and three times as many points will be deducted from Participant A's total number of points, or

- *not be implemented*, in which case no points will be deducted from either Participant B or Participant A.

Whether the decision is implemented will depend on a die roll reported by Participant B. Participant B will receive a paper cup containing a single six-sided die. Participant B will privately roll the die inside the cup, meaning that only Participant B will observe the outcome of the die roll. Participant B will then report the outcome of the die roll into the computer.

- If the reported die roll is an even number, meaning it is either **2, 4 or 6**, Participant B's decision about how many deduction points to assign to Participant A **will be implemented**. That is, Participant B's total points will be reduced by the specified number of deduction points and Participant A's total points will be reduced by three times this amount.

- If the reported die roll is an odd number, meaning it is either **1, 3 or 5**, Participant B's decision about how many deduction points to assign to Participant A **will not be implemented**. That is, Participant B will keep his or her total points and Participant A's total number of points will not be changed.

Points Calculation

The final number of points for each participant will be calculated as follows:

If Participant B's reported die roll is a 2, 4, or 6:

- Participant A's final number of points =
  + 100 (starting points)
  - Points Participant A gave to Participant B in Stage 1
  - 3 x Number of deduction points assigned to Participant A by Participant B in Stage 2 for the corresponding decision made by Participant A.

- Participant B's final number of points =
  + Points received from Participant A in Stage 1
  - Number of deduction points assigned to Participant A in Stage 2 for the corresponding decision made by Participant A

If Participant B's reported die roll is a 1, 3, or 5:

- Participant A's final number of points =
  + 100 (starting points)
  - Points Participant A gave to Participant B in Stage 1

- Participant B's final number of points =
  + Points received from Participant A in Stage 1

Your final number of points will be converted to dollars at the rate of $1 for every 10 points and added to your $5 show-up payment and $3 participation payment.

Please note that it is possible for Participant A's and/or Participant B's total number of points to be negative. In this case, the points will be deducted from the participation payment. Additional deduction points that would result in net losses greater than the $3 participation payment for Participant A or Participant B will not count against the earnings of either Participant A nor Participant B.

Final Results

All decisions will be made through the computer. After all Stages 1 and 2 have finished, a results screen will reveal the following information:

- Participant A's decision of how many points to give to Participant B in Stage 1.
- Participant B's choice of deduction points in Stage 2.
- Participant B's reported die roll in Stage 2.
- Your final number of points.
- Your total earnings for the experiment.

Are there any questions about the instructions? If you have a question, please raise your hand and wait for the experimenter.

To make sure that everyone understands the instructions, we will now proceed to some questions about the instructions, which you will answer on the computer. You may refer to these printed instructions if you need to in order to answer the questions.

Appendix C: Instructions (3rd Party Verified)

General Instructions

This is an experiment in decision-making.

In addition to a $5 show-up payment and $3 participation payment, you will be paid any money you accumulate from the experiment that will be described to you in a moment. You will be paid privately, in cash, at the conclusion of the experiment. The exact amount you receive will be determined during the experiment and will depend on your decisions and the decisions of others. *If you have any questions during the experiment, please raise your hand and wait for an experimenter to come to you. Please do not talk, exclaim, or try to communicate with other participants during the experiment.* Participants intentionally violating these rules will be asked to leave the experiment and will not be paid their earnings from the experiment.

During the experiment, payoffs will be expressed in points, not dollars. At the end of the experiment, points will be converted to dollars at the rate of $1 for every 10 points.

In this experiment, there are three types of participants: Participants A, Participants B, and Participants C. The computer will randomly and anonymously match participants into groups of three with one participant of each type. You will never be informed of the identity of the two people with whom you are matched, either during or after the experiment. Similarly, the two other participants in your group will never be informed of your identity. You will be in the same three-person group for the entire experiment.

Today's experiment consists of two stages. Any money that you earn from the experiment will be added to your $5 show-up payment and $3 participation payment. At the conclusion of the second stage, the experiment will end and you will receive your earnings in cash.

Stage 1 Instructions

At the beginning of this stage, the participants will receive the following number of points:

- Participant A: 100 points
- Participant B: 0 points
- Participant C: 50 points

In this stage, only Participant A will make a decision. Participant A will decide how many of the 100 points to give to Participant B. Participant A can give Participant B between 0 and 50 points in increments of 10 points. That is, Participant A can choose to give 0, 10, 20, 30, 40, or 50 points to Participant B.

> For example, if Participant A gives 40 points to Participant B, Participant A will have 60 points (100 – 40) and Participant B will have 40 points (0 + 40) at the end of this stage. If Participant A gives 10 points to Participant B, Participant A will have 90 points (100 – 10) and Participant B will have 10 points (0 + 10) at the end of this stage. If Participant A gives 0 points, Participant A will have 100 points and Participant B will have 0 points at the end of this stage.

Stage 2 Instructions

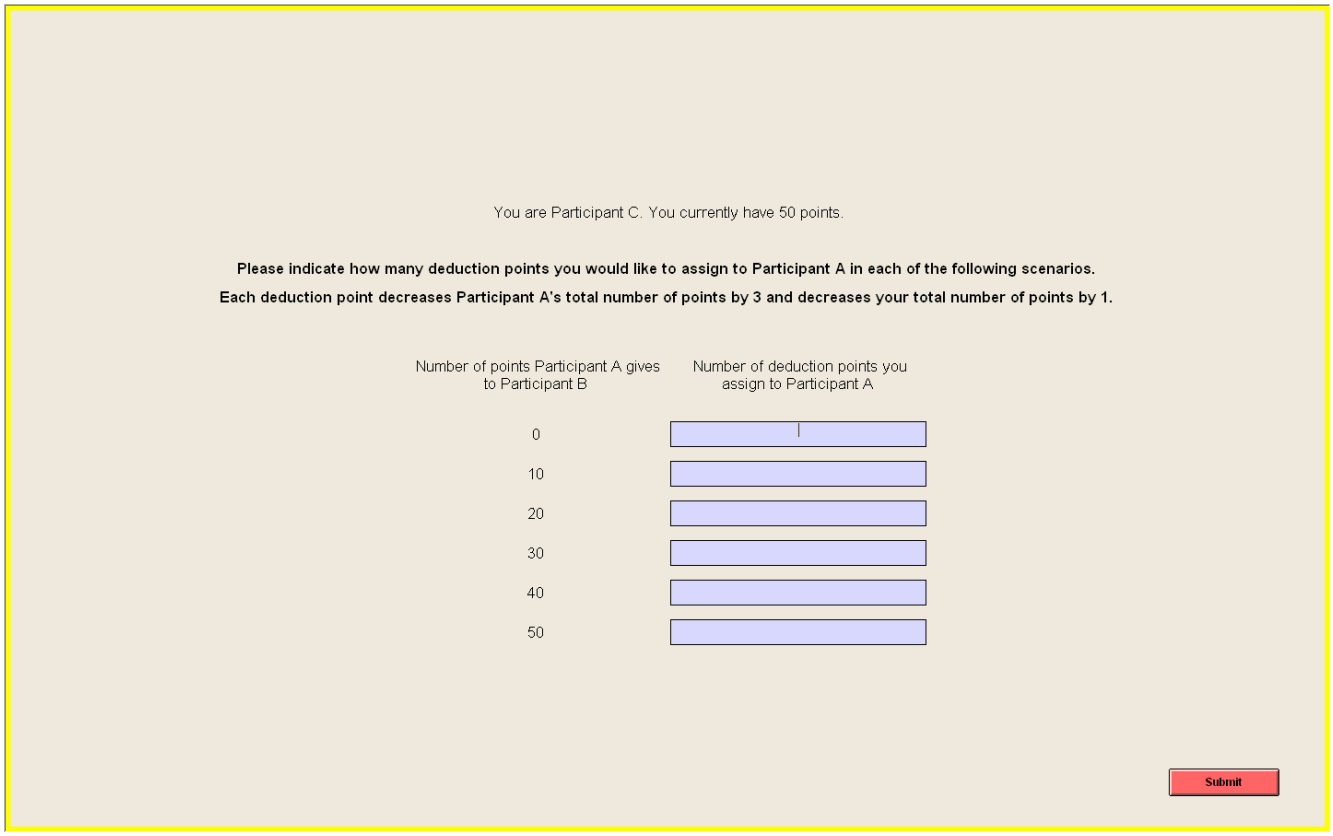
In this stage, only Participant C will make decisions.

Assigning deduction points

For each possible amount of points that Participant A gave to Participant B in Stage 1, Participant C will decide how many deduction points he or she would like to assign to Participant A. Participant C will make these decisions before knowing Participant A's actual choice in Stage 1. Participant C can assign between 0 and 50 deduction points to Participant A. For each deduction point that Participant C assigns to Participant A, Participant A's total number of points will be reduced by three and Participant C's total number of points will be reduced by one.

> For example, if Participant C assigns 2 deduction points, Participant A's total points will be reduced by 6 points (2x3) and Participant C's total points will be reduced by 2 points. If Participant C assigns 19 deduction points, Participant A's total points will be reduced by 57 points (19x3) and Participant C's total points will be reduced by 19 points.

Participant C will specify how many deduction points to assign to Participant A, for every possible amount that Participant A could give to Participant B in Stage 1. These decisions will be made on a screen like the one shown below.



Implementing the assignment of deduction points

After Participant C specifies the number of deduction points he or she would like to assign to Participant A for each possible decision by Participant A, a die roll will be used to determine whether or not Participant C's decision is actually implemented, and whether any deduction points are actually assigned to Participant A.

Participant C's decision will either

- *be implemented*, in which case the deduction points specified by Participant C, for the amount given by Participant A to Participant B, will be deducted from Participant C's total number of points and three times as many points will be deducted from Participant A's total number of points, or

- *not be implemented*, in which case no points will be deducted from either Participant C or Participant A.

Whether the decision is implemented will depend on a die roll by Participant C. Participant C will receive a paper cup containing a single six-sided die. The experimenter will observe Participant C roll the die and the experimenter will then enter the outcome of the die roll into the computer.

- If the die roll is an even number, meaning it is either **2, 4 or 6**, Participant C's decision about how many deduction points to assign to Participant A **will be implemented**. That is, Participant C's 50 points will be reduced by the specified number of deduction points and Participant A's total points will be reduced by three times this amount.

- If the die roll is an odd number, meaning it is either **1, 3 or 5**, Participant C's decision about how many deduction points to assign to Participant A **will not be implemented**. That is, Participant C will keep his or her entire 50 points and Participant A's total number of points will not be changed.

Points Calculation

The final number of points for each participant will be calculated as follows:

If Participant C's die roll is a 2, 4, or 6:

- Participant A's final number of points =
  + 100 (starting points)
  - Points Participant A gave to Participant B in Stage 1
  - 3 x Number of deduction points assigned to Participant A by Participant C in Stage 2 for the corresponding decision made by Participant A.

- Participant B's final number of points =
  + Points received from Participant A in Stage 1

- Participant C's final number of points =
  + 50 (starting points)
  - Number of deduction points assigned to Participant A in Stage 2 for the corresponding decision made by Participant A

If Participant C's die roll is a 1, 3, or 5:

- Participant A's final number of points =
  + 100 (starting points)
  - Points Participant A gave to Participant B in Stage 1

- Participant B's final number of points =
  + Points received from Participant A in Stage 1

- Participant C's final number of points =
  + 50 (starting points)

Your final number of points will be converted to dollars at the rate of $1 for every 10 points and added to your $5 show-up payment and $3 participation payment.

Please note that it is possible for Participant A's total number of points to be negative. In this case, the points will be deducted from Participant A's participation payment. Additional deduction points that would result in net losses greater than the $3 participation payment for Participant A will not count against the earnings of either Participant A nor Participant C.

Final Results

All decisions will be made through the computer. After all Stages 1 and 2 have finished, a results screen will reveal the following information:

- Participant A's decision of how many points to give to Participant B in Stage 1.
- Participant C's choice of deduction points in Stage 2.
- Participant C's die roll in Stage 2.
- Your final number of points.
- Your total earnings for the experiment.

Are there any questions about the instructions? If you have a question, please raise your hand and wait for the experimenter.

To make sure that everyone understands the instructions, we will now proceed to some questions about the instructions, which you will answer on the computer. You may refer to these printed instructions if you need to in order to answer the questions.

Appendix D.

Sections 3.1 and 3.2 reported treatment comparisons on the frequency of subjects who chose to punish at least one of the possible actions by the first party. To check the robustness of our results, we also examined the frequency of reported even rolls among those who punish at least once, twice, and so on. The results for each are reported below. All the p-values are for binomial tests of whether the reported number of even rolls is statistically different from an unbiased random process.

Treatment	3P Self Report	2P Self Report	3P Verified
Total number of subjects	58	61	55
• Number of punishers who punished in at least 2 scenarios.	16	32	15
Number (frequency) of even rolls	4 (25%)	22 (69%)	-
p-value	0.077	0.050	
• Number of punishers who punished in at least 3 scenarios.	15	30	14
Number (frequency) of even rolls	4 (27%)	21 (70%)	-
p-value	0.118	0.043	
• Number of punishers who punished in at least 4 scenarios.	13	25	11
Number (frequency) of even rolls	2 (15%)	18 (72%)	
p-value	0.022	0.043	
• Number of punishers who punished in at least 5 scenarios.	10	18	6
Number (frequency) of even rolls	2 (20%)	13 (72%)	
p-value	0.109	0.096	

We also considered the number of instances (i.e. different amounts given by Participant A) under which a potential punisher stated an intention to punish and test differences by treatment with

Kolmogorov-Smirnov tests. We find that the distribution of number of instances of stated punishment significantly differs between the 2P Self Report and 3P Self Report conditions ($p=0.03$), between the 2P Self Report and 3P Verified conditions ($p=0.03$), but not between the 3P Self Report and 3P Verified conditions ($p>0.99$). Repeating the above comparison while excluding subjects who reported odd die rolls gives the same results (p-values of 0.03, 0.03, and 0.99, respectively).